Interakce člověk–počítač v přirozeném jazyce (ICP)

LS 2013 — Teorie skrytých Markovových modelů

Tino Haderlein, Elmar Nöth

Katedra informatiky a výpočetní techniky (KIV)
Západočeská univerzita v Plzni

Lehrstuhl für Mustererkennung (LME)
Friedrich-Alexander-Universität Erlangen-Nürnberg

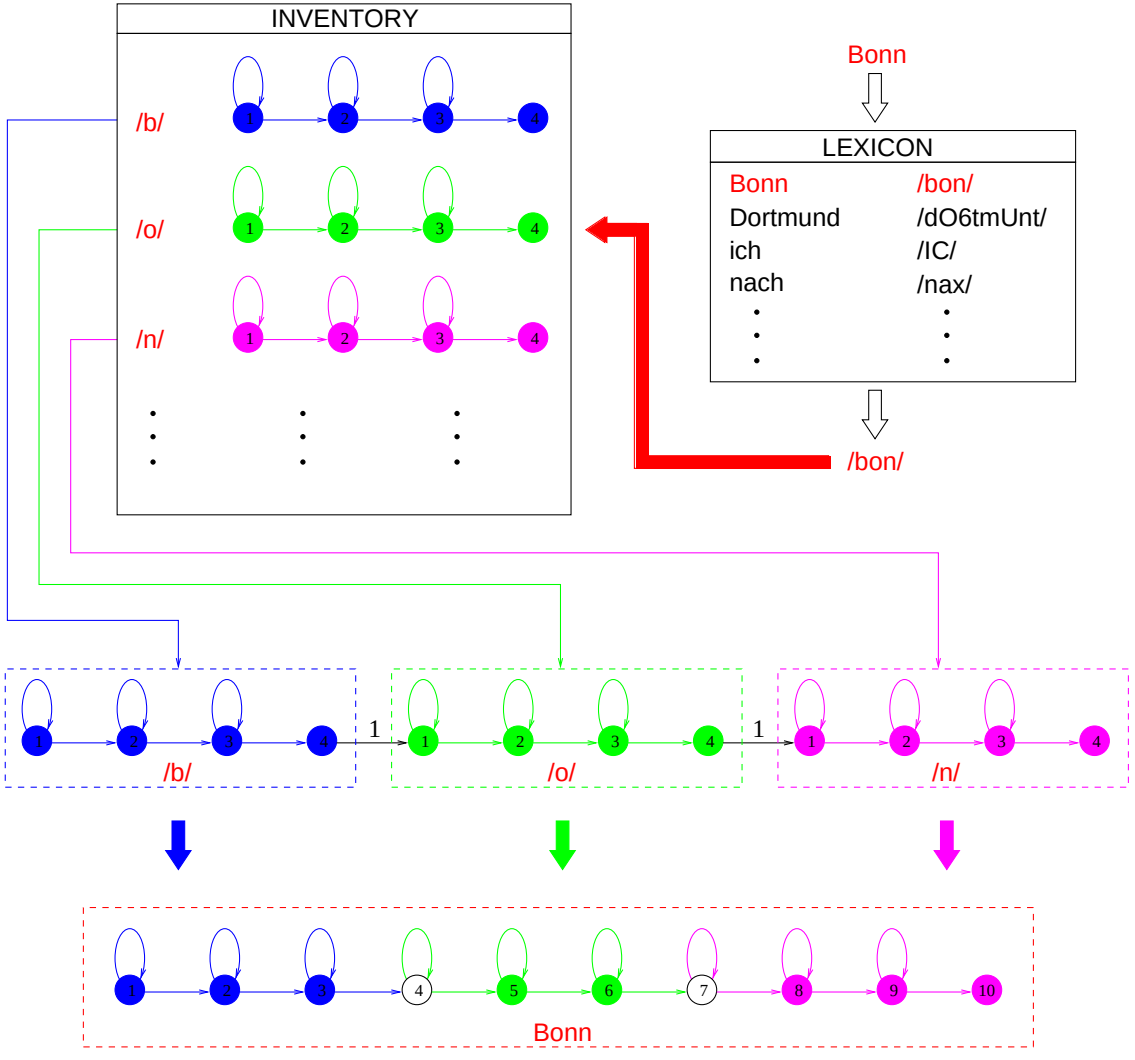
Inventář modelů

- Hodně slov ve slovní zásobě na rozpoznávání
  $\rightarrow$ modelování každého slova vlastním HMM není možné:
  - Je příliš hodně slov, příliš málo trénovacích dat pro slovo.
  - Možná pro několik slov ani nejsou příklady promluvy.
  - Přidání nových slov k aplikaci (např. titul nového filmu v informačním systému pro kina) není moc dobře možné: nejsou příklady promluvy, popř. mají různé varianty výslovnosti být modelovány?
- Příklad: Slovní zásoba má 1000 slov.
- Pro robustní odhad modelu slova potřebujeme nejméně10–100 příkladů výslovnosti jednoho mluvčího.
- Když je průměrná délka slova asi 0,4 sekund (němčina), je minimální potřeba tréninkových dat $0{,}4 \cdot 10 \cdot 1000 = 4000$ sekund, tj. asi 1–10 hodin dat jednoho mluvčího.

Inventář modelů

- Řešení: „analýza syntézou“
  - Každé slovo lze skládat sekvencí částí slov (angl. subword units).
  - Je jen omezená množina (inventář) částí slov.
  - Každý prvek inventáře se modeluje vlastním HMM.
  - Modely slov se skládají sestavováním (konkatenací) modelů kratších jednotek.
  - Tím je mnohem méně HMM než slov, je proto dost tréninkových dat pro každý HMM.
  - Nové slovo do slovníku: Žádná tréninková data pro slova nejsou nutná, jen posloupnost částí slov.

Inventář HMM pro jednu část slov

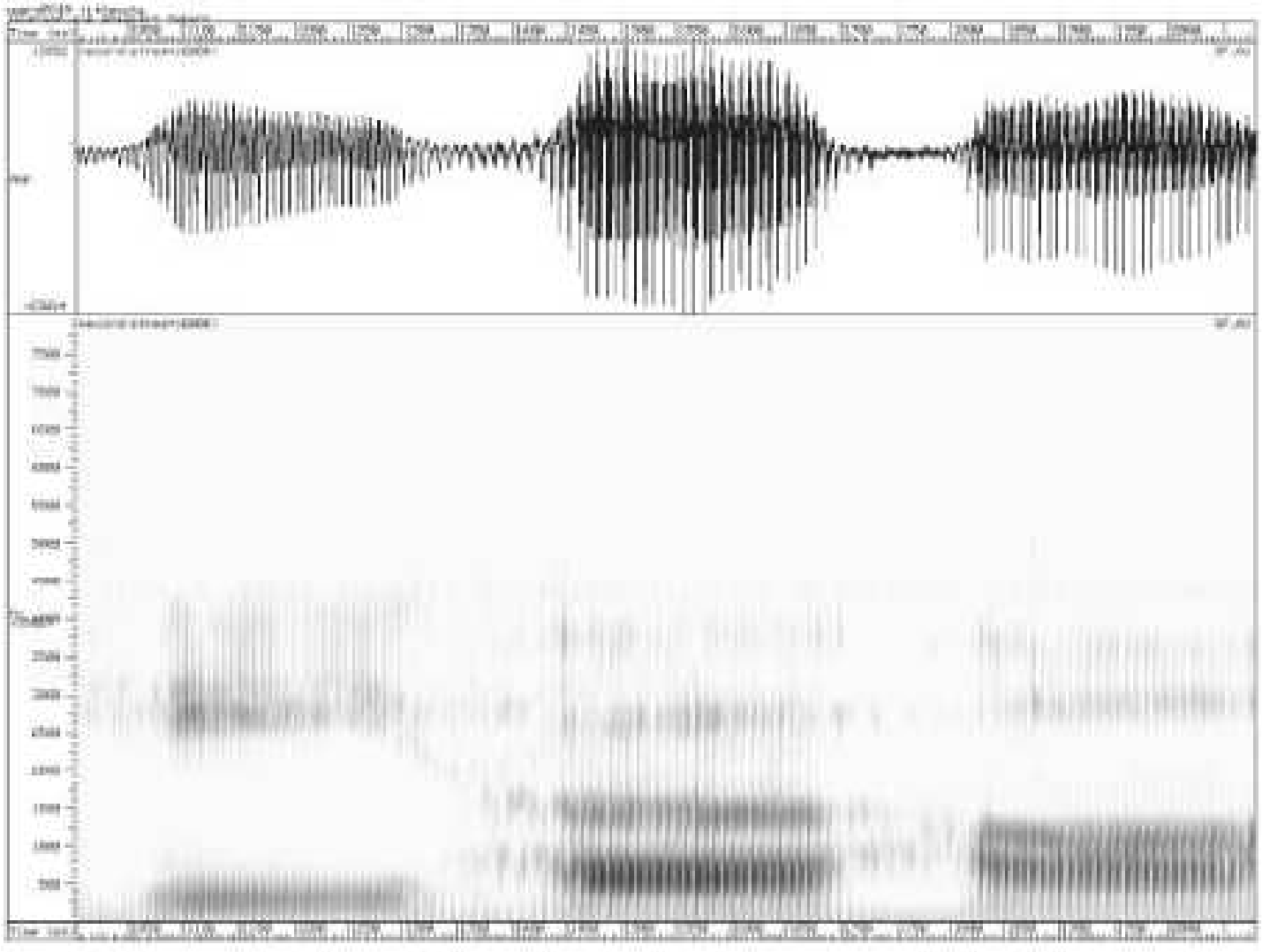


Výběr částí slov

- Dříve rozpoznávání řeči užívalo hodně různých částí slov, např. hlásky, slabiky, ...
- Rozhodnutí pro určité části slov se má orientovat podle těchto kritérií:
  - **Přesnost:** Prvky inventáře by měly být specifické vzhledem k výslovnosti a s ostatními by se měly překrývat nebo k variacím obměňovat jen málo.
  - **Trénovatelnost popř. robustnost:** Inventář by mělo mít málo prvků, aby byl dostatek dat pro každý HMM.
  - **Modularita:** V ideálním případě lze každé slovo jazyka skládat konečním inventářem částí slov.
  - **Transfer popř. zobecnitelnost:** Pro syntézu modelů nových slov by mělo být ke konstrukci úvod, který vyžaduje žádné náročné znalosti znalce (např. fonetika).

Výběr částí slov

koartikulace: (anglicky) hi h3: hA:

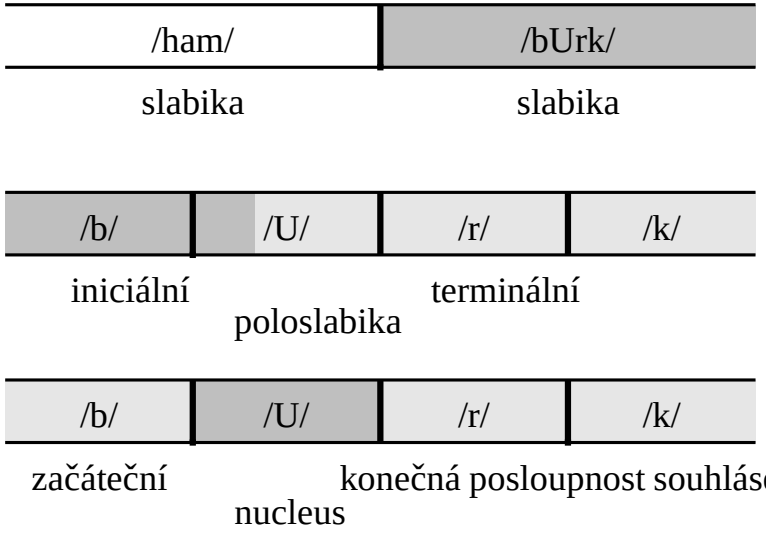


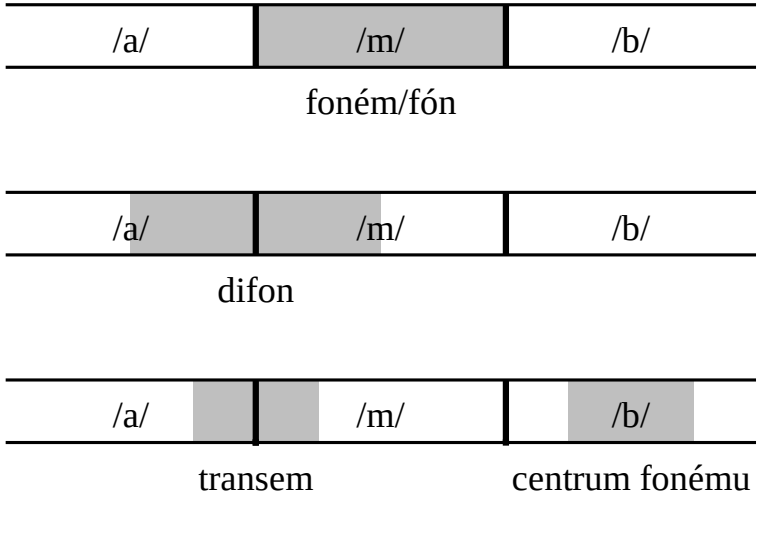
Výběr částí slov

- Příklad:
  - Hlásky jsou robustně trénovatelné, protože je jich málo.
  - Hlásky mají velkou modularitu a dobré vlastnosti, co se týká transferu (nutná je jen fonetická transkripce nového slova).
  - Přesnost modelů hlásek je velmi špatná, protože hlásky mluvčí vyslovují ve závislosti na kontextu velmi různě ($\rightarrow$ koartikulace).
  - Slova jsou velmi precizní, mají ale špatné vlastnosti robustnosti a modularity.
- Rozlišuje se dva typy částí slov:
  kontextově nezávislé a kontextově závislé části slov

Kontextově nezávislé části slov

- Část slov se buď orientuje na fonologickou strukturu slova, anebo je řízená daty stanovena automatickou metodou.
- Pro každý prvek inventáře části slov se trénuje přesně jeden HMM.
- Dříve se užívalo hodně různých kontextově nezávislých částí slov.
- Část slov se vybere tak, že kontextově podmíněná variace je malá, popř. že kontext, který má vliv na variaci, je už součástí části slov: **zmrazení** kontextu





Kontextově nezávislé části slov: příklady

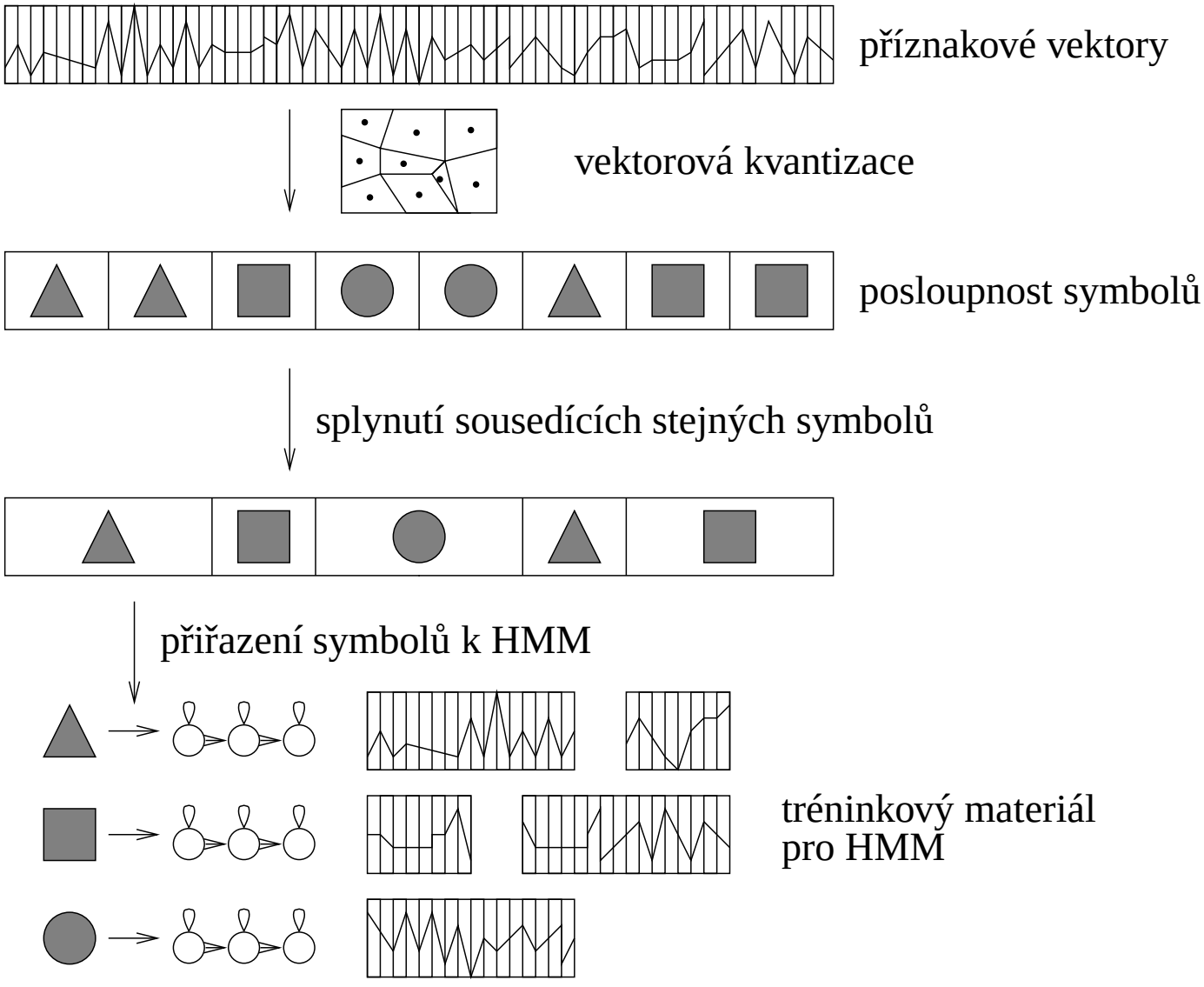
- **fóny**
  mezi 40 a 200 univerzálních jednotek
  velmi modulární a přesnější než fonémy (jeden foném $\rightarrow$ $>$1 fón)
  fonetická notace už není jednoznačná
- **slabiky**
  20 000 angl., 100 japon. slabik, ve slovanských jazycích kolem 2500–3000
  v němčině 50 000 slabik možných, ale objevuje se jen malá část; špatná modularita, např. pro 1000 slov $>$ 1000 slabik
  výborná rozlišovací ostrost, koartikulace uvnitř vyslovené slabiky
- **poloslabiky**
  rozdělení slabiky do iniciální a terminální poloslabiky
  $\rightarrow$ menší inventář, ale stejná přesnost
  800 ($\approx$1000) iniciálních a 2560 ($\approx$1000) terminálních německých (anglických) poloslabik

Řízené daty, kontextově nezávislé části slov: fenony

- Stanovení části slov by se mělo stát empiricky s množinou dat tak, že uvedená kritéria jakosti jsou optimalizovaná.
- Výsledná část slov se jmenuje fenon.
- Různé postupy stanovení inventáře fenonů a fenonického základního tvaru $F = f_1, ..., f_m$ slova, např.:
  1. Trénuj vektorový kvantizér na datech.
  2. Přiřaď ke každému příznakovému vektoru symbol kódové knihy kvantizéru.
  3. Spoj po sobě následující symboly, které jsou stejné.
  4. Každý symbol kódové knihy odpovídá fenonickému symbolu f a je reprezentován pomocí HMM $\lambda(f)$.
- Když jsou různé příklady výslovnosti $\boldsymbol{X}^1, ..., \boldsymbol{X}^n$ jednoho slova, vybereme optimální fenonický základní tvar $F^\star$ tak, že maximalizuje pravděpodobnost vytvářet všechny n příklady:

$$F^\star = \text{argmax}_F \prod_{i=1}^{n} P(\boldsymbol{X}^i | \lambda(F))$$

Fenony


příznakové vektory
vektorová kvantizace
posloupnost symbolů
splynutí sousedících stejných symbolů
přiřazení symbolů k HMM
tréninkový materiál
pro HMM

Fenony

Výhoda:

- Vektorová kvantizace tvoří modulární, precizní, robustně trénovatelné jednotky.

Nevýhoda:

- Pro každé nové slovo je potřebný příklad výslovnosti $\rightarrow$ fenony jsou málo užíváné.

Kontextově závislé části slov

Problém kontextově nezávislých částí slov:

- Když kontext, který ovlivňuje jednotku, má být součástí jednotky, jednotka musí být poměrně dlouhá.
- Robustní odhad delších jednotek je ale složitý.

Postup u kontextově závislých částí slov:

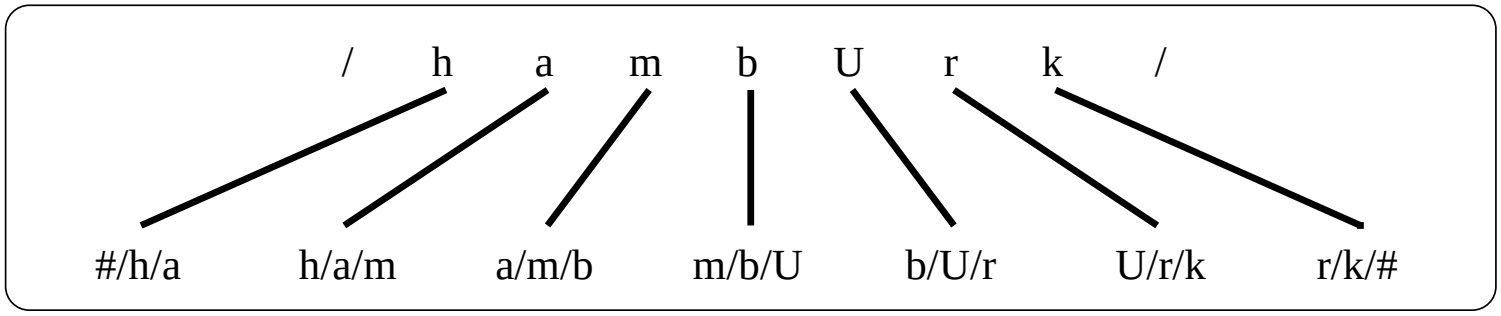
- Jednotka modelovaná pomocí HMM je krátká.
- Ale pro každý kontext, ve kterém se jednotka nachází, se trénuje vlastní HMM.

$\rightarrow$ robustně odhadnuté a přesto precizní modely

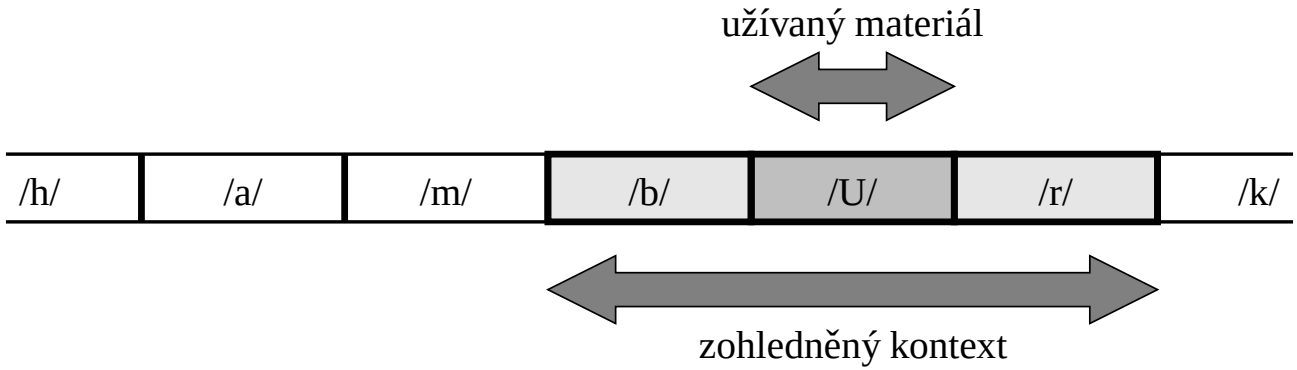
Kontextově závislé části slov

- Většina kontextově závislých částí slov se zakládá na fonémech v různých kontextech (alofonech).
- Předpokládáme, že je realizace fonému závislá jenom na sousedních hláskách.
  - trifony: Zohledňuje se pouze levá a pravá sousední hláska
    /r/ v /hambUrk/ $\longrightarrow$ U/r/k .
  - pravé/levé bifony: Zohledňuje se pouze pravá/levá sousední hláska
    r $\longrightarrow$ /r/k
    r $\longrightarrow$ U/r/ .
  - monofony: Žádný hláskový kontext není zohledňován –
    r $\longrightarrow$ /r/ .
- Hranici slova reprezentuje vlastní symbol #.
- důležité: bifon není difon!

Trifony



- Příklad: Trifon U/r/k charakterizuje v němčině centralizovanou hlásku [6] s **regresivní asimilací** vzhledem k **palatální plozivě** /k/.
- HMM trifonu b/U/r je trénován jen s příznakovými vektory hlásky U, ale ne s každými /U/ v tréninkových datech, ale pouze se všemi /U/, kde levý kontext je /b/ a pravý kontext je /r/.



Trifony

- Většina trifonů se neobjevuje v tréninkových datech.
- Trifon-HMM má ale přesně stejnou topologii jako bi- a monofon-HMM stejného fonemu.
- Strategie „recyklace“ menších jednotek:
  - žádný trifon $\Rightarrow$ vezmi pravý bifon
  - žádný pravý bifon $\Rightarrow$ vezmi levý bifon
  - žádný levý bifon $\Rightarrow$ vezmi monofon

$$\text{b/U/r} \rightarrow \text{/U/r} \rightarrow \text{b/U/} \rightarrow \text{/U/}$$

- Monofony se mohou užívat i pro inicializaci modelů bi- a trifonů.

Trifonová interpolace

- Trifony jsou velmi precizní, ale příslušné HMM často nejsou odhadnuté robustně.
- Monofony a bifony jsou méně precizní, ale odhad je robustnější.
- Lineární interpolace parametrů HMM mono-, bi- a trifonů sdružují výhody těchto jednotek.
- Interpolační váhy se mohou stanovit heuristicky (proporcionálně na četnost v datech).
- Interpolační váhy se mohou optimalizovat algoritmem EM na separátní množině dat.

Generalizované trifony

problém: vzácné trifony $\Rightarrow$ parametry HMM jsou statisticky špatně odhadované

- řešení: splynutí trifonů, které patří ke stejnému fonému, a podobných kontextů
- cíl: sdružování různých sousedních fonémů, které ovlivňují artikulaci centrálního fonému stejným způsobem
- postup: sdružování/generalizace buď zakládající se na fonetickém vědomí anebo řízené daty

Generalizované trifony

- např. systém s 5 třídami pro levé/pravé sousední fonémy (původně pro němčinu)

jádrový foném je samohláska	jádrový foném je souhláska
hranice slova, aspirace nebo /h/	hranice slova nebo aspirace
labiální souhlásky	palatální samohlásky nebo /j/
dentální, alveolární nebo palatální souhlásky	zaokrouhlené samohlásky nebo /w/
velární souhlásky	nezaokrouhlené samohlásky
samohlásky	souhlásky

→ pro jádrový foném je možných jen $5 \cdot 5 = 25$ trifonů

Řízená daty generalizace trifonů

např. pomocí phoneme environment clustering (PEC):

1. Začni s robustně odhadnutými modely monofonů.
2. Udělej několik binárních rozdělení vzorové oblasti:
   Každé půlení zvýší počet (generalizovaných) trifonů.
3. Opakuj do té chvíle, než budou trifonové modely robustně odhadnuté a precizní.

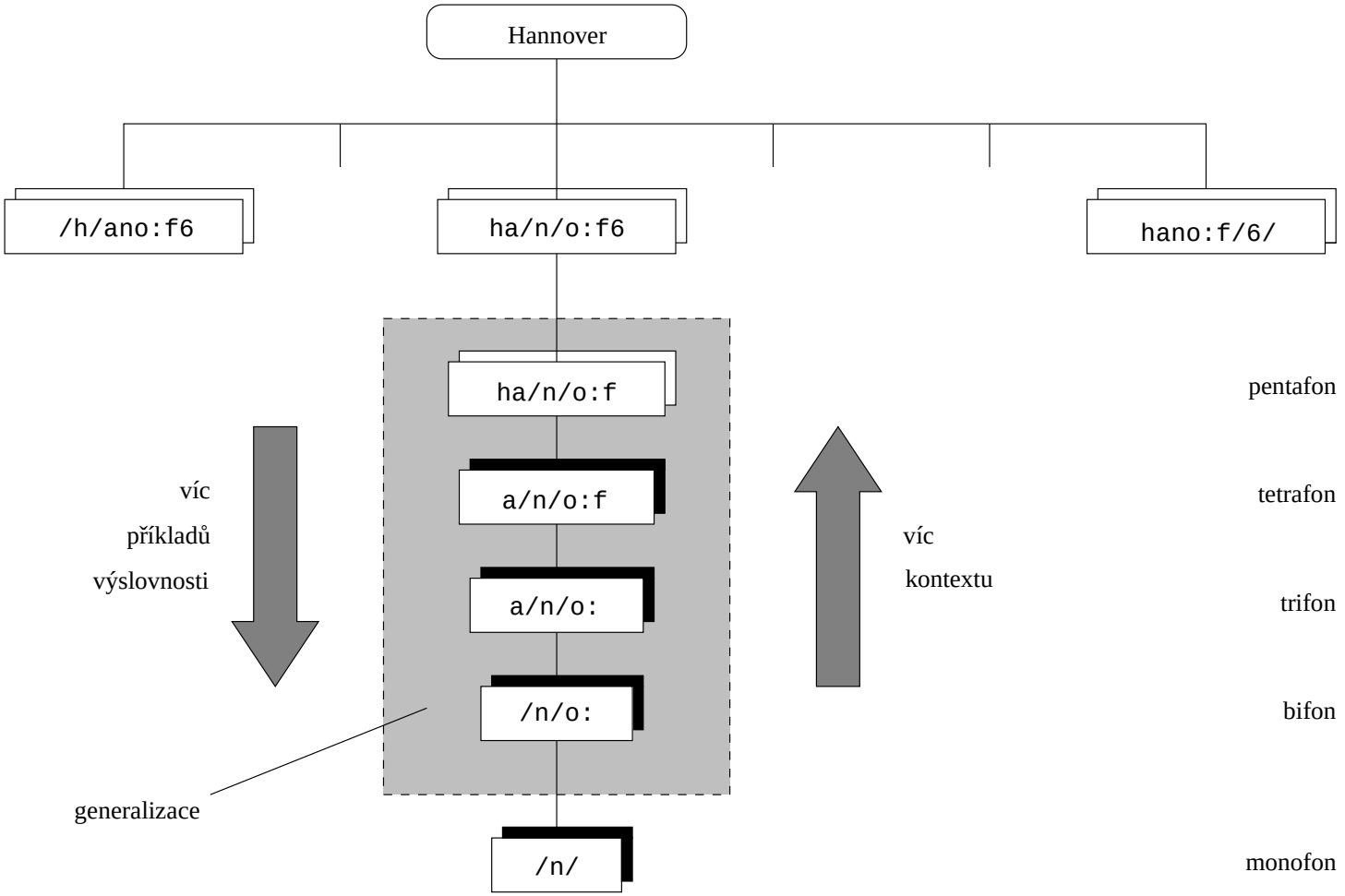
Binární dělení vzorové oblasti např. pomocí CART (classification and regression tree):

- Stanov určitou zásobu otázek, např. „Je pravý sousední foném přední samohláska?“
  Možné odpovědi pro každý trifon jsou „ano“ a „ne“.
- Při každém binárním dělení vybereme otázku, která při dělení optimalizuje entropii (maximální zisk informací).

Polyfony

- I když jazykové jednotky s dlouhým kontextem (tetra-, penta-, hexa-, heptafony, . . .) jsou poměrně velmi vzácné, přesto je v každé množině dat těch jednotek hodně, které se objevují často.
- Je **možné** je modelovat (lze robustně odhadnout HMM).
- Je ale i **nutné** je modelovat (silný vliv nadbytečnosti a zkreslení)
  $\rightarrow$ fonémy v **libovolně širokém** pravém/levém kontextu
  $\rightarrow$ foném závislý na celém slově.
- Zhrubnutí **vyváženým odřezáváním**; vpravo a vlevo, zvenku dovnitř
- Dodatečně: **označování přízvuku** a **hranic** (fráze, slovo, morfém, slabika)
- Trénuj HMM pro polyfon, když se objevuje v datech víckrát než dolní práh, např. 50krát.

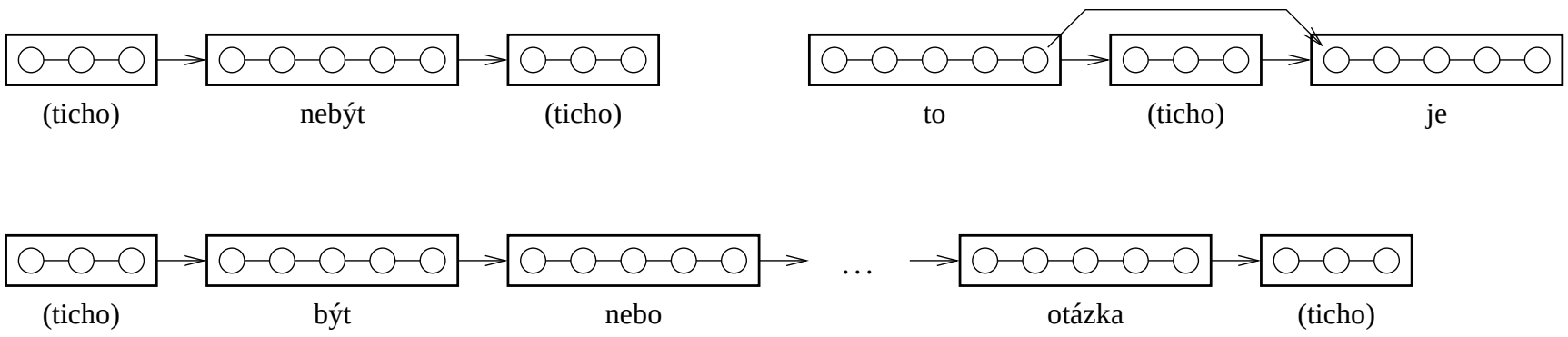
Polyfony



- hierarchická reprezentace, ekonomický postup
- Jeden vzor může sloužit na trénink **vícero HMM**.
- Polyfony využívají tréninkovou množinu **nejlépe**.

Trénink HMM a části slov

- V tréninkových datech se běžně říkají celé věty a ne jednotlivá slova nebo dokonce části slov.
- V tréninku proto spojíme jednoduché modely do komplexního HMM pro celou větu:
  - Části slov a slova spojíme s pravděpodobností přechodu 1.
  - Mezi slovy se vloží doplňkové HMM ticha.
  - Baum-Welchův algoritmus zařídí korektní přiřazení mezi vzorem a modelem.



Neexistující slova a neznámá slova

V reálné aplikaci se vždycky objevují

- promluvy spontánní řeči,
- zvuky okolí,
- extralingvistické promluvy,
- slova mimo slovní zásobu.

Rozpoznávač přiřazuje ke každé neznámé promluvě nejpodobnější slovo své slovní zásoby $\rightarrow$ chyby.
Řešení: dodatečné modely pro

- nonverbální fenomény jako chrchlání, kašlání, smích, zvuky dýchání nebo mlaskání
- pochybování: vyplněné a nevyplněné pauzy
- rušivé zvuky: např. tútání telefonu, klepání
- neznámá slova

Modelace slovních hranic

- koartikulace **nejen uvnitř** slova:
  např. zkreslení krátkých funkčních slov:

  ten muž /tEn/ + /mUZ/ $\rightarrow$ /tEmuZ/
- trifonové modely překrývající slova: crossword triphones
- **kombinatorika kombinace hlásek** mezi dvěma slovy
  $\Rightarrow$ **ztrojnásobení** počtu různých trifonů (od 1800 na 5500)
- nové trifony mezi slovy jsou **vzácné**
- jen nové trifony, které se objevují nejméně 30krát
  $\Rightarrow$ počet stoupá o méně než 10
- generalizace/interpolace disponibilního **inventáře modelů**
  na trifony mezi slovy **modelu dekódování**
- **rozpoznávací fáze:** sousední slova a hláskový kontext neznámé!